08901-F

Research on Probabilistic Information Processing

Final Report

Principal Investigator:
WARD EDWARDS

July 1973

Ames Research Center
National Aeronautics and Space Administration
Grant No. NGL-23-005-171
Moffett Field, California 94035

Final Report

RESEARCH ON

PROBABILISTIC INFORMATION PROCESSING

July 1973

Prepared for:

Ames Research Center
National Aeronautics and Space Administration
Code DRU/M/S 201-5
Moffett Field, California 94035

Attention: R. H. Sutton,
Staff Assistant University Affairs

Prepared by:

Engineering Psychology Laboratory
Institute of Science and Technology
The University of Michigan

Prepared under:

Grant NGL-23-005-171
Principal Investigator: Ward Edwards

Introduction

On 1 December 1965 the University of Michigan submitted to NASA a proposal for research on probabilistic information processing. Work actually began on 15 June 1967; funding was $80,000 on that date; $80,000 on 1 July 1968; $80,000 (step-funding) on 1 July 1969; and $75,000 (extension of step-funding) on 1 July 1969. The step-funding ran out on 30 June 1972. This is the final report of the program of research. It covers work done between 15 June 1967 and 30 June 1972. The first section contains a general summary of the changes in our thinking that occurred as a result of our basic and applied research. The second section presents a detailed account of what we promised to do, as contained in our proposals, and what came out of those promises. The third section details those products delivered that had not been promised. The final section contains a documentation of the articles and reports emanating from this research effort along with abstracts summarizing research results.

One comment is perhaps worth making about the program as a whole before proceeding to the body of the report. Originally, our intention had been to perform work of direct significance to the function of the Mission Operations Control Room in connection with manned space flight missions. We are somewhat unclear about the degree to which our work in fact had such directly NASA-relevant impact. But the ideas generated during the program have had meaningful impact on other non-NASA U.S. government activities. The general idea of probabilistic information processing, brought from a gleam in the eye to a feasible technology in the course of this program, is now routinely applied by the Defense Intelligence

Agency and other organizations in the U.S. intelligence community. Multi-attribute utility measurement, using technology developed in this program, is now in use as a program planning tool in the Office of Child Development, DHEW, and is likely to spread to other program planning and evaluation activities. While these are the most important impacts on government programs, other less direct ones can easily be traced.

This is one "applications-oriented program" that really did produce applications.

Section I

In 1966 EPL was concerned with two specific questions. One was, "what is the locus of conservatism?" The second was, "is PIP a better information processing system than existing systems?" We asked these questions as part of our ongoing attempt to identify those aspects of the environment that affect the way people process information and to determine those properties of the individual that account for his inability to extract the proper amount of information from data. We could ask these questions because we had an appropriate methodology including specifiable data generating processes that could be implemented.

Our concern with understanding both the environment and the characteristics of the information processor has been motivated by the realization that the more we know about information processing, the better our chance of improving it. Thus, even while we did primarily basic research within controlled laboratory environments we never felt that the issues we addressed should be investigated for only academic reasons. If our research product could not lead toward better information processing systems or better decisions, then we should be doing something else. In 1966, we felt that we were ready to test the PIP ideas in real world situations. Consequently, we came to NASA searching for contexts where we could implement PIP on-line.

However, how should we test out PIP? We expected PIP to do at least as well as any other system at all times and better than other systems in situations

having an intermediate level of uncertainty. If there is very little uncertainty, almost any system should be correct. If there is great uncertainty, there may be no basis for any system to reach an appropriate conclusion. Consequently, we needed to find situations with some uncertainty, but not too much. PIP should also be better in situations where a great deal of data must be processed including those data of minimal diagnostic value that are otherwise frequently overlooked. Therefore, we hoped to find situations where there was a lot of slightly diagnostic data.

Before we could test PIP to see if it was better than its alternatives, we needed to define the concept "better." We have spent considerable time and effort on this non-trivial problem, and have concluded that many operational definitions are appropriate. If the data-generating process is well-defined, then probabilities calculated by means of it can be compared directly with those calculated from human estimates. If the data-generating process is not known but the correct hypothesis is known, then probability estimates can be evaluated (over an ensemble of such instances) by means of what are called proper scoring rules. If neither the data-generating process nor the correct hypothesis is known, then analogy from laboratory experiments with known data-generating processes suggests the idea that more extreme probability distributions are more nearly correct than less extreme ones. (Obviously this last idea must be handled with care. The n-th power of a probability distribution, renormalized, is also interpretable as a probability distribution--more extreme than the first power, if n is greater than 1. Distributions can be too extreme as well as too flat, though extremeness is clearly preferable to flatness whenever the data justify it. Analogy from experiments with known data-generating processes suggests that PIP is unlikely to be too extreme.) From a different point of view, a "better" system is one which the users find more acceptable.

For each of these definitions of "better" statistical problems arise in measuring deviations from optimality or in comparing one system with another, and in aggregating such measures over situations, systems, subjects, and data within a situation. We have worked out approaches to these problems for each of the definitions of "better" we considered.

Before PIP could be put to any test at NASA or anywhere else, implementation problems had to be solved. This need dictated a new orientation of our basic research program. Response modes, display and training questions, and the problems of how to handle conditionally dependent data and nonstationary hypothesis situations were studied. Experiments were conducted in which subjects used either a verbal response mode or recorded their responses on logarithmically spaced scales. An experiment was run comparing the effects on LR estimation of the feedback from a display showing the posterior probability distribution implied by a given set of LRs (likelihood ratios) for a particular datum to LR estimation where there was no visual display as feedback. We carried out experiments in which subjects estimated LRs and revised their odds for both conditionally dependent and conditionally independent data. Other experiments compared odds estimation in both cascaded and noncascaded hypothesis situations. We ran an experiment to determine whether people respond appropriately to nonstationary environments.

As the same time that this basic research program was undertaken we were searching for contexts within Mission Control where we could implement and test PIP on-line. However, it became apparent that MSC was not the place to test these ideas for two main reasons. First of all, mission controllers do not deal with

much irreducible uncertainty and consequently all rational systems should lead to the same conclusions. Secondly, mission controllers deal with a very rich hypothesis space. In situations like this where so many hypotheses are being considered simultaneously, the cost in time and effort of getting all the necessary inputs would probably exceed any potential gain of this system over the existing one.

While we did not find that MSC offered a good test of PIP, we came to believe that MSC might provide a good opportunity to test other aspects of decision analysis technology. It appeared to us that the formulation of Mission Rules might be helped by means of decision analysis, since such rules are essentially pre-specified decisions.

Therefore we set out to study other aspects of decision analysis, especially the measurement of value or utility. In particular we wanted a method of validating utility judgments comparable to the known-data-generating-process case for probabilities. Knowing the "right answer" would enable us to use departure from optimality as the criterion by which to evaluate performance. We have not yet succeeded in implementing a known utility approach. We found that existing environmental models that we examined were too complex. Furthermore, the laboratory models that we devised could not be realistic, camouflaged and evaluateable simultaneously. Therefore we tried a different approach. We tested experimentally both in the laboratory and in the field the feasibility of the Weighted Linear Average (WLA) Model. We used as a criterion the predictability of choices after extended learning based on aggregated judgments from choices based on decomposed judgments. In undertaking both the laboratory and the on-line studies we found

ourselves faced with new response mode, training and display problems that needed solutions. We considered different methods of dealing with interdimensional incomparability. One method is to compare every dimension with a single dimension such as money or time saved. Another method involves constructing appropriate lotteries and determining relative preferences for these lotteries. We used several procedures, including direct-rating techniques with normalized and unnormalized quantities, to elicit assessments of the relative weights for the different dimensions of a multi-attributed object. We tried different schemes for defining the end points of the scales. In one scheme the end points of the individual scales were marked with ambiguous descriptions such as very important vs. not important. In other studies the end points were designated by the extremes of the dimension that had some realistic probability of occurring. This technology however is still in the very beginning stages of development.

Our laboratory and field activities have led to some conclusions, and also to some rather important non-NASA applications.

One important conclusion is that probabilistic information processing systems can be implemented and do work. We initially were much concerned about response modes, nonstationarity, violations of conditional independence, and training. But all of these problems seem soluable, and in fact have been solved in applications of Bayesian information processing to intelligence system data.

Our conclusions about conservatism complicate the picture somewhat. The Wheeler thesis, a NASA-sponsored product, shows that in laboratory situations misaggregation lies behind the phenomenon of conservatism. This finding seems to bring the abstract laboratory question to a natural termination, and we do not

anticipate doing any more such studies. But what does this mean for real-world probabilistic information processing?

Two practical facts of life intervene at this point. One is that most real world information processors are inexperienced with expressing uncertainties as probabilities. To introduce the rather indirect reasoning implicit in a PIP system to such people all at once is too much. First they must come to regard uncertainty as naturally and properly measured by probability--and this is in itself a major and prolonged educational problem. After that, it remains the case that, as our NASA research shows, feedback from a Bayesian system to the estimators of quantities to go into it will seriously degrade the system--and yet such feedback is imperative in practice if the operators are to be willing to use and trust the system. For these reasons, we anticipate that the full-dress probabilistic information processing systems originally envisaged will develop only gradually, emerging from difficulties with direct estimation of probabilities. Indeed, some response modes so blur the distinction between PIP systems and direct estimates of posterior probabilities that it becomes rather difficult to decide which system is which.

Are such systems conservative? We don't know, but suspect so. Does that conservatism impair their performance? Probably. Are such effects eventually reflected in sub=optimal decisions? We don't know. But such questions must finally be looked at in real systems, not only in the laboratory, though relevant laboratory work remains to be done.

The next major topic in research on decision analysis seems to us to be the structuring of the decision problem. The formal structure of decision analysis is well-defined; various appropriate response modes exist; elicitation and training procedures have been studied at least in part. A look into the shape of decision-theoretical maxima has been very enlightening to us. Unless the problem has unusual asymmetries in payoffs or probabilities, substantial departures from optimal strategy produce only small percentage reductions in expected value of the act chosen. This general insensitivity implies that response modes and training procedures are less important than we used to think they are. But decision analyses are extremely sensitive to the basic framework used. Addition of one more act, consideration of one more state, concern about one more dimension of value--all of these can utterly change an analysis. How can the basic framework of a decision analysis be studied as a scientific topic? A NASA-sponsored doctoral dissertation on the structuring of hierarchical inference systems has provided us with a small piece of the answer by showing how such systems can be structured, and how the structure influences the system's behavior. But the surface of this topic is only scratched.

Section II

The research program that was completed at EPL under NASA sponsorship can be arbitrarily divided into the following four categories: Research in the MSC Setting, Laboratory research on PIP, Research on multi-attribute utilities, and Small scale Bayesian research.

1. Research in the MSC Setting

In the 1966 proposal it was stated that the goal of our on-site research efforts would be to prepare a document showing in detail how PIP could be applied in the NASA context. However, before this could be done, EPL personnel would have to familiarize themselves with NASA command problems and settings by reading documents and by visiting NASA facilities.

EPL personnel visited MSC and read scores of documents. The result of this effort was a realization that PIP per se was not a useful system for MSC mainly because mission controllers do not deal with very much irreducible uncertainty and secondly because they do deal with a very large set of hypotheses. The first reason means that PIP does not have much to offer over existing systems. The implication of the second reason is that using PIP would be too costly in time and effort. Consequently, our emphasis shifted away from an exploration of the usefulness of PIP to an exploration of the usefulness of decision analysis. The 1968 and 1969 proposals reflected this change.

In 1968 and 1969 we proposed to explore the possibility that explicit estimation of probabilities and values is useful as a basis for writing mission

rules. More specifically, we suggested that whenever the formulation of a mission rule was both difficult and important enough to warrant a formal procedure, the decision analysis, exploiting human judgment, might be a reasonable course to pursue. The specific plan of action had three phases. During Phase 1 EPL personnel would become sufficiently familiar with the rules for some particular mission to select a few rules that met our criteria. Phase 2 would consist of preparing and running simulations that exercise these chosen rules. During this phase explicit probability and value judgments would be collected. In Phase 3 new mission rules where appropriate would be written. These new rules would be more consistent with the judged values and probabilities obtained in Phase 2 than the old rules were. Then these new mission rules would be tried out in simulated missions. Another purpose of either Phase 2 or 3 would be to see what effects simulation experience would have on the assessed value and probability judgments.

In order to explore the extent to which the values and probabilities that enter into mission rules can be made explicit test simulations were formulated and run according to our specifications (Edwards, 1968). These studies led to several conclusions. First, nothing in the exercises denied our premise that every decision, including those made on-line by space vehicle controllers, depends on subjective answers to the questions: What's at stake and what are the odds. Second, while irreducible uncertainty plays a remarkably small role in on-line control of space flight, value judgments play an extremely major role. Finally, formal decision theory with its explicit use of cardinal probability and value judgments can make a contribution to resolving difficult cases. Moreover, the

decision theoretic approach has a great deal to offer a more highly automated mission control system.

We had at least three objectives at MSC--to test our ideas in a real world setting, to teach our techniques to NASA personnel, and to find relevant problems where our technology would make a definite contribution. We were only partially successful. We were not able to persuade MSC personnel of the value of using decision analysis to construct mission rules.

One of the major ramifications of our interaction with MSC was an increasing awareness of the importance of developing a utility measurement technology as quickly as possible. In this spirit we proposed to NASA in 1970 to develop such a methodology for use in solving the problems of selection, scheduling and rescheduling of experiments. This proposal was not supported.

2. Laboratory research on PIP

Based on prior laboratory research on PIP, we believed that PIP showed promise of being a viable diagnostic system with advantages over existing systems. But PIP still needed the test of real world applications. However, in order to implement a system of this kind, for test or other purposes, practical, how-to-do-it procedures need to be established. Therefore, in 1966 we proposed the first of a series of studies on how to best tool up a PIP system. This experiment proposed to investigate the effects of different kinds of response modes and displays.

The study and the data analyses were completed and the results were disseminated. These results are being incorporated into a review article now in progress (Goodman).

In the 1968 proposal three additional studies along this same line were proposed. The purpose of the first was to explore the question of whether group discussion combined with a requirement for consensus led to essentially the same results as averaging pre-group individual likelihood ratio (LR) assessments. This study was done by Barbara Goodman as her doctoral dissertation research. It has been completed and the journal article published (Goodman, 1972).

The purpose of the second experiment was to explore the differences, if any, in LR assessment between those Ss who assess P(D|H) and are shown the LRs implied by their estimates and those Ss who assess LRs directly and are shown the P(D|H) values implied by their assessments. This study was not done. The most similar study we did was a pilot study designed to see whether Ss would change their LRs or their odds when the two sets of estimates led to different posterior odds for the same sequences of data. The problem in doing the original study as proposed and in our failure to draw any conclusions from the pilot study that was done has been training. If an experimenter uses untrained subjects, then he can generalize about the information processing ability or lack of it in the general population. However, if one is concerned about designing information processing systems using trained operators, then the results of these studies have their limitations. What does a well-trained PIP operator know? What can he do? What experience has he had?

In order to have an operable PIP system we need to resolve the question about what to do about the absence of conditional independence in the data set. The beginnings of an answer to this question was proposed as the third experiment in the 1968 proposal. Its specific purpose was to investigate two particular procedures. One procedure consists of collecting dependent data together and

treating each set as a single datum. The other procedure consists of estimating LRs that are conditional on preceding data.

This experiment was completed and the journal article published (Domas and Peterson, 1972).

Keeping with the spirit of "let's implement a PIP system", the 1969 proposal suggested experiments in two of the problem areas previously opened up--training and conditional dependence of the data set.

A transfer of training experiment was proposed in which subjects would be trained to estimate LRs in a system having a specifiable data generating process partially known to the subjects. It would probably be two formal distributions differing only in their mean values. These same subjects would then be asked to estimate LRs in systems having DGPs both similar and dissimilar to the DGP in the training situation.

This experiment was not done.

Beginning with the 1969 proposal, research on conditional dependence was being viewed as part of the larger problem of cascaded or multi-stage inference. There were two questions posed on this topic in the 1969 proposal. One was, "how can a probability distribution over a set of data be treated as a datum?" The other was, "is conservatism a phenomenon in a cascaded as well as a single stage inference system, assuming the system has either a binomial or a normal DGP?"

The first question remains unanswered. Three experiments on the second question have been completed and the respective journal articles will appear in a forthcoming issue of Organizational Behavior and Human Performance devoted to this topic (Youssef and Peterson, a; Youssef and Peterson, b; and Gettys, Kelly and Peterson.)

In the 1970 proposal, research on PIP was concentrated on the topic of multi-stage inference. Two main experimental tasks were suggested. One was to uncover an explanation that predicts radicalism in multi-stage inference systems and conservatism in single stage systems and then adequately tests out this hypothesis. The other task was to determine what changes are required in the technology of PIP to handle datum unreliability and hierarchical organization of information processing.

The experiments on this topic have resulted in a much greater understanding of this problem than we had anticipated. One doctoral dissertation has been produced (Kelly, 1972) and two journal articles (Snapper and Fryback, 1971; Gettys, Kelly and Peterson, in press) have been written.

3. Research on multi-attribute utility measurement (MAUM)

The 1968 proposal first called for EPL to do research on this topic. What we outlined was a small scale laboratory study entitled "Research on the measurement of value." We proposed to design an experiment that would explore different methods for applying the Weighted Linear Average (WLA) Model. To evaluate the different approaches we realized that our first task would be to define an appropriate experimental situation having stimuli where an objective standard of correctness of the composite exists, such as prices for used cars. Thus, in 1968 we began our search for such an experimental situation. We are still searching.

In 1969 the general purpose of our MAUM program was to study how men make value judgments, how they can be helped to make them better and how such judgments could be used as inputs to decision-making systems. Specifically we proposed to

investigate the WLA model as a model for MAUM. The first step was to define a dependent variable; preferably, the deviations of subjects' judgments from optimality. However, this requires that you have an official standard of judgment, that is, a model of what we sometimes call "God's utility function." It must have the nature of a many-to-one transformation. It must be complex enough so that no amount of experience with its outputs would make it transparent and obvious to the subjects. Furthermore, it must be reasonable and intuitive so that subjects don't have too much unlearning to do. The next step would be to train subjects to have intuitions that roughly parallel "God's utilities." Given that we have been able to find a suitable model and have elaborately trained our subjects, we proposed as the next step a set of experiments that would enable us to begin investigating the WLA model. These experiments would examine two necessary procedural questions--how do you elicit single dimension value judgments from subjects and how do you elicit importance judgments for each dimension.

A "God's utility" model incorporating all of the features that we want was not developed. However, one study was done under NASA sponsorship that did have many of these features and did investigate several methods for eliciting value judgments and importance weights (von Winterfeldt, 1971).

In the 1970 proposal a multi-attribute utility measurement procedure was proposed to handle the problems of selecting, scheduling and rescheduling experiments for manned space flights. This procedure consisted of the following steps: identifying the dimensions of value relevant to each experiment; locating each experiment on each dimension; rescaling the dimension; judging of the

importance of each dimension, applying the WLA equation and finally testing. In order to develop the appropriate methodology to implement this procedure we proposed to conduct work on the four kinds of experiments that would be needed enroute. These were experiments on response modes, abstract validation experiments at the individual level, less abstract validation experiments at the individual level, and finally explorations within the real world context.

Edwards (1970) summarizes our thinking in this area.

4. Small scale Bayesian research

In 1966 a single small scale Bayesian research study was proposed. It was a laboratory experiment on the topic of the locus of conservatism comparing individuals who estimated either LRs or odds in either a cumulative or non-cumulative mode.

This experiment was done using the pick-up stick data generating process for the first time. A journal article is in preparation (Wheeler and Edwards).

In 1968 two major categories of small scale studies were proposed in addition to the category on value measurement which we have already included under the multi-attribute utility measurement section. One category proposed continued research on conservatism investigating such topics as the relative contribution of number of data and diagnosticity of each datum to conservatism and the confusion concerning primacy and recency effects. One experiment that was specifically proposed was a study to test the hypothesis that it is only with a 50-50 prior distribution that the first datum of a sequence is properly processed. Our motivation was a desire to investigate the frequent finding that estimation of the diagnostic value of the first datum was the Bayesian prescribed value in spite of the fact that this assessment required an aggregation step,

that is, combining the prior odds with the assessing of the impact of the first datum.

This research package was the basis for the studies done by DuCharme as his doctoral dissertation research. An article summarizing this set of experiments has been published (DuCharme, 1970).

The other category of research proposed in 1968 was the topic of information purchase. Experiments were outlined concentrating on the flat maximum problem, i.e., the problem that the expected value function is flat relative to the independent variable being manipulated.

Several experiments on this topic were run using subjects' direct estimates of the value of information as the dependent variable. A journal article summarizing the results of these studies has been published (Wendt, 1969).

The 1969 proposal outlined subsequent studies under the same categories of research detailed in 1968. More work was proposed on the topic of the locus of conservatism. More specifically, more research was suggested to tie down more firmly the conclusion that misaggregation is the primary cause of conservatism.

The several experiments that comprise our last word (at least for awhile) on this topic were done by Wheeler as her doctoral dissertation research (Wheeler, 1972). A journal article summarizing these studies is in preparation.

The information purchase issue that we proposed to investigate was the change, if any, in subject's optimality as a function of the sharpness of the function relating EV to the amount of information purchased.

This research was completed and a journal article is in preparation (Saltzman and Peterson).

There were no additional small scale Bayesian research studies detailed in the 1970 proposal. By this time we had made a commitment to spend our energies on studying multi-attribute utilities.

Section III

Any large scale research effort extending over a several year period that allows for any flexibility will produce research products not promised in any proposal. Our program has produced several of these "bonuses."

In January, 1969, Edwards prepared an updating of his 1964 bibliography on publications on human decision processes. This new work (Edwards, 1969) contains 1393 references.

In implementing a PIP system, one of the major problems is how to handle a nonstationary environment. Chinnis and Peterson asked the basic question, "can people discern when an environment has changed?" The results of this study have been published in a journal article (Chinnis and Peterson, 1970).

One variation of the information purchase model that we have investigated is specifically relevant to the question of how people make the tradeoff between speed and accuracy in tasks in which time costs money. An extensive series of experiments have been conducted on this topic. The results are published in two journal articles, Swensson and Edwards (1971) and Swensson (1972). This latter publication describes the research that Swensson completed for his doctoral dissertation.

Wendt conducted several information purchase experiments testing the hypothesis that it might be the discreteness of the variables of the decision matrix that caused the conservatism found in previous studies. The results of these studies were presented in a paper given at the 3rd Research Conference on

Subjective Probability, Utility, and Decision Making, London, on September 7, 8, and 9, 1971 (Wendt, 1971).

Wallsten looked at the locus of conservatism question within the framework of conjoint measurement theory and then related this approach to the Bayesian scheme and the notions of misperception, misaggregation, and response bias. This development and the results of an experiment to test some of the implications of this development have been published in a journal article (Wallsten, 1972).

Saltzman ran an experiment both to determine whether subjects are conservative information processors when making probability assessments over a continuum and to investigate alternative response modes for credible intervals. This study was completed, the data analysis finished and a draft of the journal article prepared (Saltzman).

Section IV

1. Edwards, W. Controller decisions in manned space flight. In Applications of Research on Human Decision Making, Proceedings of a symposium on Application of Research on Human Decision Making, 1968, Washington, D. C.: NASA Scientific and Technical Information Division, NASA-SP-209, 1970, 93-106.

This paper describes the results of two simulation studies run for mission controllers. Four controllers were interviewed in detail after the simulations had been run. They were asked to estimate the values and odds bearing on each decision and, in the case in which information processing was called for, an appropriate likelihood ratio for the datum to be processed. When (as usually happens) some answers were inconsistent with others, the controllers were invited to revise any or all answers in the direction of greater consistency. The result was that all controllers did achieve essentially consistent sets of estimates, and that in all cases the estimates predicted the decision that was actually made. In other words, a computer, given the same value and probability estimates that the controller had made, could have made the same decisions.

2. Edwards, W., Phillips, L. D., Hays, W. L. & Goodman, B. C. Probabilistic information processing systems: Design and evaluation. IEEE Trans. Syst. Sci. Cybernetics, 1968, 3, 248-265.

A Probabilistic Information Processing System (PIP) uses men and machines in a novel way to perform diagnostic information processing. Men estimate likelihood ratios for each datum and each pair of hypotheses under consideration (or a sufficient subset of these pairs). A computer aggregates these estimates by means of Bayes's theorem of probability theory into a posterior distribution that reflects the impact of all available data on all hypotheses being considered. Such a system circumvents human conservatism in information processing, the inability of men to aggregate information in such a way as to modify their opinions as much as the available data justify. It also fragments the job of evaluating diagnostic information into small, separable tasks. The posterior distributions that are a PIP's output may be used as a guide to human decision making, or may be combined with a payoff matrix in order to make decisions by means of the principle of maximizing expected value.

A large simulation-type experiment compared PIP with three other information processing systems in a simulated strategic war setting of the 1970's. The difference between PIP and its competitors was that in PIP the information was aggregated by computer, while in the other three systems, the operators aggregated the information in their heads. PIP processed the information dramatically more efficiently than did any competitor. Data that would lead

PIP to give 99:1 odds in favor of a hypothesis led the next best system to give 4 1/2:1 odds.

An auxiliary experiment showed that if PIP operators are allowed to know the current state of the system's opinions about the hypotheses it is considering, they perform less effectively than if they do not have this information.

This paper reports work done before the NASA program began; only its preparation was supported by NASA.

3. Wendt, D. Value of information for decisions. J. Math. Psychol., 1969, 6, 430-443.

Information that will reduce the risk of a decision may be costly in time, effort, or money. The maximum amount that should be invested in the information--its fair cost--depends upon payoffs, the diagnosticity of the data source, and prior probabilities of the hypotheses. These are the independent variables of this experiment. Subjects estimated the fair costs by means of the Marschak bidding procedure. The subjects' bids changed in the direction appropriate to each of the three independent variables, but not enough to be optimal.

4. Edwards, W. Man-machine systems for policy mediation and intellectual control. Talk given at the Fourth Annual NASA-University Conference on Manual Control, Ann Arbor, Michigan, March 22, 1968.

No further publication of this speech is planned.

5. Edwards, W. A bibliography of research on behavioral decision processes to 1968. Human Performance Center Memorandum Report No. 7, January 1969.

This report lists 1393 references to publications on behavioral decision processes to 1968. It is an alphabetical listing by author and consequently is not grouped according to any topic categorization scheme.

6. Swensson, R. G. The elusive trade-off: speed versus accuracy in visual discrimination tasks. Perception & Psychophysics, 1972, 12, 16-32.

Theoretical models for choice reaction time and discrimination under time pressure must account for Ss' ability to trade accuracy for increased speed. The fast guess model views these tradeoffs as different mixtures of "all-or-none" strategies, while incremental models assume they reflect different degrees of thoroughness in processing the stimulus. Three experiments sought tradeoffs for difficult visual discriminations, using explicit payoffs to control and manipulate pressures for speed and accuracy. Although guessing was pervasive, the simple fast guess model could be rejected; Experiments II and III obtained tradeoffs even when fast guesses were purged from Ss' data. Tradeoff functions fit by several formulations revealed: a) slower rates of increase in accuracy for more similar stimuli, and b) substantial "dead times" (80 - 100 msec slower than detection times) before discrimination responses could exceed chance accuracy. Errors were sometimes faster and sometimes slower than correct responses (depending on S's speed-accuracy trade); the latter effect may reflect a ceiling on S's achievable accuracy. This paper ended with a discussion that examines the implications of the results for models of discrimination under time pressure; it suggests modifications in present models, focusing on the random walk model, and describes an alternative "deadline" model.

7. Chinnis, J. R., Jr. & Peterson, C. R. Nonstationary processes and conservative inference. J. exp. Psychol., 1970, 84, 248-251.

The experiment tested the hypothesis that people are conservative processors of fallible information because they treat stationary data-generating processes as if they were nonstationary, i.e., subject to change from time to time. The Ss made inferences from fallible data when the population from which the data were sampled could change during the sampling process. Performance on this task was compared with performance on a similar, but stationary task. The Ss behaved differently in the two situations, appropriately assuming zero probability of change only in the stationary task. In addition, the pattern of conservatism in the two tasks requires rejection of the hypothesis that conservatism is due to inappropriate assumptions of nonstationarity.

8. DuCharme, W. M. Response bias explanation of conservative human inference. J. exp. Psychol., 1970, 85, 66-74.

Conservative human inference has been attributed to misperception or misaggregation of data, but it may be caused by response biases. In the present experiments, Ss revised odds estimates about which one of two normal distribution data generators was being sampled. An analysis of special sequences and a plot of revised odds against theoretical odds in Exp. I showed a bias in Ss' response functions. They revised odds optimally only over a range of ± 1.0 log odds. When E set different levels of prior odds, the response functions shifted so that the optimal range centered around the set prior odds. A second experiment showed that the biased functions remained invariate over changes in data generator familiarity and diagnosticity. Of the several explanations offered for these response functions, an odds bias seems the most likely. Whatever the cause of the bias, Ss neither misaggregated nor misperceived data within their optimal range.

9. Goodman, B. C. Action selection and likelihood ratio estimates by individuals and groups. Organizational Behavior and Human Performance, 1972, 7, 121-141.

This study investigated the shifts between individual and group performance in a choice dilemma, a gambling, and a Bayesian likelihood ratio estimation task. Twenty seven male subjects performed each task alone. Six four-man leaderless groups were formed and repeated the each task. Three subjects performed the task alone a second time. Finally, all 27 subjects repeated each task again alone. The choice dilemma task decisions reproduced previously found patterns of shifts. Groups preferred higher variance gambles than did the average of pregroup individuals. The post-group likelihood ratio estimates of 22 of the 24 test subjects resembled their group's estimates more closely than they resembled their own pregroup estimates. Both group and individual correlations between measures of performance in all three tasks were low.

10. Snapper, K. J. & Fryback, D. G. Inferences based on unreliable reports. J. exp. Psychol., 1971, 87, 401-404.

Inferences may be based on direct observation of events or on reports from indirect sources about the occurrence of events. A direct observation will be more diagnostic than a report if the source of the report is not completely reliable. Previous studies have investigated Ss' inferences based on either directly observed events or completely reliable reports. This study investigated Ss' inferences based on partially reliable reports. The Ss responded to reduced report reliability by using a formally inappropriate rule that led to overestimation of the diagnostic impact of a report.

11. Swensson, R. G. & Edwards, W. Response strategies in a two-choice reaction task with a continuous cost for time. J. exp. Psychol., 1971, 88, 67-81.

Each trial of a two-choice task rewarded S for a correct response but charged a cost proportional to his response time. Seven of the eight Ss in three experiment violated predictions of the random-walk model and confirmed those of the fast-guess model by using only two response strategies in all conditions. These Ss either responsed accurately or made a detection response when the stimulus appeared, accepting chance-level error rates to respond 15-20 or 45-70 msec. faster (for two different types of stimuli). Stimulus frequency and payoffs primarily determined which strategy S would adopt. Data were ambiguous for only one S equally well fit by the random-walk model and by assuming that he intermittently guessed on some proportion of trials.

12. Wendt, D. Use of information in a risky market situation. Paper presented at 3rd Research Conference on Subjective Probability, Utility, and Decision Making, London, Sept. 7, 8, 9, 1971.

A series of experiments investigated purchase decisions of Ss in the role of a retailer in a risky market situation. To aid their choice of a quantity to be bought, they had tables indicating their payoff for all possible combinations of amount bought and amount vendible, and information from a data source correlated to the amount vendible (state of nature). Results showed that Ss were sensitive to the diagnosticity of the data source but not precisely as prescribed by the normative model of expectation maximation: their decisions were too much influenced by data of low diagnostic value (radicalism), and too little influenced by highly diagnostic value (conservatism). Displaying tables of conditional probabilities of states of nature (amount vendible) given the datum (P(H|D)) led to slightly better decisions (in the sense of expectation maximation), than displaying probabilities of data given states of nature (P(D|H)), or joint probabilities (P(D&H)). Another variable studied was the shape of the conditional probability distribution (trapezoid vs. a normal approximation).

13. Domas, P. A. & Peterson, C. R. Probabilistic information processing systems: Evaluation with conditionally dependent data. Organizational Behavior and Human Performance, 1972, 7, 77-85.

Previous research on Probabilistic Information Processing (PIP) systems has used data that are conditionally independent. In the real world, data are frequently conditionally dependent, that is, given an hypothesis, the occurrence of one datum influences the likelihood of occurrence of a second datum. If the use of a PIP system is desired when the data are known to be conditionally dependent, then it is necessary to know if PIP is an appropriate system for the processing of dependent data. One experiment compared PIP with a second system POP. PIP operators gave more optimal estimates when the data were conditionally independent; however, POP estimators gave more optimal estimates when the data were conditionally dependent. A second experiment attempted unsuccessfully, to produce a modified PIP system that would give optimal estimates for both conditionally dependent and conditionally independent data. A revised technology of diagnostic information processing based on conditionally dependent data was proposed.

14. Wallsten, T. S. Conjoint-measurement framework for the study of probabilistic information processing. Psychol. Rev., 1972, 79, 245-260.

Certain assumptions are invoked, implicitly or explicitly, whenever a descriptive model of human processing of probabilistic information is built around Bayes's rule. This paper shows that the two primary assumptions are equivalent to a very general additive conjoint-measurement model. These plus an additional assumption concerning the nature of sequential effects are then proved to be equivalent to a distributive conjoint-measurement model, given a certain task restriction. This allows the three assumptions to be tested with ordinal data. It also provides a framework for the investigation of specific theoretical problems concerning the nature and determinants of composition rules and scale values. The notions of misperception, misaggregation, and response bias currently discussed in the literature may be viewed as one subset of those problems. This approach generates experiments of a sort not previously done. One of them is presented herein. The basic model was not rejected for most of the 12 subjects, and certain diagnostic properties demonstrated how it failed for the others. In addition, specific relationships concerning the scale values emerged. The paper concludes with a discussion of the theoretical prospects resulting from this development.

15. Youssef, Z. I. & Peterson, C. R. Intuitive cascaded inferences. Organizational Behavior and Human Performance, in press.

Previous research has investigated the process by which people make single-stage inferences, i.e., how they revise probability estimates about hypotheses at one level as the result of observing data at an immediately lower level. Such intuitive probability revisions usually turn out to be conservative with respect to optimal performance. The present experiments investigated the inferential process with two-stage inferences; it was necessary to cascade information from Level 1 to Level 3 in a hierarchically organized, probabilistic situation. The cascaded inferences turned out to be systematically more excessive than corresponding noncascaded inferences. The excessiveness was maintained as the data took on several different levels of diagnostic value.

16. Wheeler, G. E. Misaggregation versus response bias as explanations for conservative inference. University of Michigan PhD Thesis, 1972.

Recent research in human information processing has shown that most subjects are conservative in making inferences from evidence. Conservative inference is the process of extracting less than the optimal amount of certainty from data. Several hypotheses have been proposed to explain conservatism, namely misperception, misaggregation, and response bias.

Bayes's theorem provides the mathematically correct way of incorporating data into an uncertain situation. The misperception hypothesis asserts that people incorrectly perceive the impact of single data items, but use Bayes's theorem to combine data; they do the right arithmetic with the wrong numbers. The misaggregation hypothesis is that people correctly perceive the impact of a single datum, but do not combine data correctly; they do the wrong arithmetic with the right numbers. The response bias hypothesis, which has several variations, asserts that over the range of values with which people are familiar, they both perceive and aggregate data correctly, but outside that range do neither very well. The experiments reported in this paper were designed to test the various hypotheses, and in particular to compare misaggregation and response bias predictions. Earlier research had indicated that misperception, while certainly present under many conditions, was not the primary locus of conservatism.

In the present experiments, subjects made odds estimates about which of two previously defined populations was being sampled. The two populations were approximately normal distributions. In Experiment I, d' (the separation between the means of the distributions) was varied; subjects responded to three different pairs of data generating populations. When subjects were required to revise their odds over a sequence of data, they were conservative, regardless of the size of the theoretically correct odds. When they estimated odds for individual data, they were veridical, even when outside the range that the response bias hypothesis would predict as veridical. Subjects were sensitive to the diagnosticity of the data generating populations, and responded appropriately to the different d' values. The results were descriptive both of median estimates and individual subjects' responses.

In Experiment II, the way of displaying the data generating populations was varied, as was the way in which sequences were constructed. The results indicated that neither of these parameters affected the way in which subjects responded. In all conditions, subjects continued to display conservatism when required to aggregate data, and to show veridicality when considering individual data. As in Experiment I, the results were descriptive of individuals' responses as well as of median estimates.

The results of this study indicate that although both misperception and response biases exist, conservative inference may be primarily attributed to misaggregation.

17. Kelly, C. W., III. Application of Bayesian procedures to hierarchical inferences. University of Michigan PhD thesis, 1972.

The state of knowledge required to solve an inductive inference problem will often be partitioned so that it is difficult or impossible to directly estimate likelihoods linking observable events and target variables, i.e., data and hypotheses. For example, one state of knowledge, ξ_1, might describe distributions involving a random variable, d_i, and a second state of knowledge, ξ_2, might describe distributions for a random variable, h; but neither state is sufficient to calculate the joint distribution $\{d_i,h\}$. In situations like this, states of knowledge can be combined to yield $\{d_i,h\}$ if one or more variables e^j, called intermediate or explanatory variables, can be found such that $\{d_i\ e^j. \ . \ .\} \in \xi_1$ and $\{e^j \ . \ . \ .h\} \in \xi_2$.

Inferences which must incorporate one or more intermediate or explanatory variables are called hierarchical, cascaded, or multi-stage inferences. The work described herein is concerned with formulating a normative model for hierarchical inference, investigating the theoretical properties of the model, and applying the model to the solution of real world inference problems.

The formulation begins by describing a series of hierarchical inferences which arise in various situations. Structural or organizational properties common to all these inferences are identified. It is shown that these structural properties can be captured by modeling a hierarchical inference as an upper

semi-lattice the nodes of which represent hypotheses, data and intermediate variables. The edges of the graph describe statistical correlations which link the data, intermediate variables and hypotheses. Drawing upon the Bayesian concept of a probability as an orderly opinion, conditional probabilities are associated with the edges of the graph and an algorithm is developed for solving the general hierarchical problem.

Theoretical results are derived for a number of special cases of the general model. Various analyses show how the impact of a datum entering a hierarchical system is affected by the conditional probabilities, the number of levels in the hierarchy, and symmetry conditions. The performance of the optimal model is compared with that of two sub-optimal models in which one or more or the inference makers in a hierarchy is assumed to operate in an information reduction mode. In some cases, this does not result in as serious a loss of performance as had been expected. In this vein, the performance of natural hierarchical systems is also contrasted with that of the optimal model.

Several case studies are presented in which the model is used to solve real world diagnosis problems. Since most of the problems are concerned with unique or nearly unique events, considerable emphasis is placed on procedures for eliciting probabilities from relevant experts. A new procedure based on the use of second-order probability density functions is developed and techniques used to generate the inference tree or map the problem structure are described. The case studies show that the use of hierarchical decomposition is an aid in communicating and substantiating conclusions, encourages dialog and establishes rules for debate.